

Model Reduction for Nonlinear Control Systems using Kernel Subspace Methods

Jake Bouvrie

Department of Mathematics

Duke University

Durham, NC 27708 USA

jvb@math.duke.edu

Boumediene Hamzi

Department of Mathematics

Imperial College London

London SW7 2AZ U.K.

b.hamzi@imperial.ac.uk

Abstract

We introduce a data-driven order reduction method for nonlinear control systems, drawing on recent progress in machine learning and statistical dimensionality reduction. The method rests on the assumption that the nonlinear system behaves linearly when lifted into a high (or infinite) dimensional feature space where balanced truncation may be carried out implicitly. This leads to a nonlinear reduction map which can be combined with a representation of the system belonging to a reproducing kernel Hilbert space to give a closed, reduced order dynamical system which captures the essential input-output characteristics of the original model. Empirical simulations illustrating the approach are also provided.

I. INTRODUCTION

Model reduction of controlled dynamical systems has been a long standing, and as yet, unsettled challenge in control theory. The benefits are clear: a low dimensional approximation of a high dimensional system can be manipulated with a simpler controller, and can be simulated at lower computational cost. A complex, high dimensional system may even be replaced by a simpler model all together leading to significant cost savings, as in circuit design, while the “important variables” of a system might shed light on underlying physical or biological processes. Reduction of linear dynamical systems has been treated with some success to date. As we describe in more detail below, model reduction in the linear case proceeds by reducing the dimension of the system with an eye towards preserving its essential input-output behavior, a notion directly related to “balancing” observability and controllability of the system. The nonlinear picture, however, is considerably more involved.

In this paper we propose a scheme for balanced model-order reduction of general, nonlinear control systems. A key, and to our knowledge, novel point of departure from the literature on nonlinear model reduction is that our approach marries approximation and dimensionality reduction methods known to the machine learning and statistics communities with existing ideas in linear and nonlinear control. In particular, we apply a method similar to kernel PCA as well as function learning in Reproducing Kernel Hilbert Spaces (RKHS) to the problem of balanced model reduction. Working in RKHS provides a convenient, general functional-analytical framework for theoretical understanding as well as a ready source of existing results and error estimates. The approach presented here is also strongly empirical, in that observability and controllability, and in some cases the dynamics of the nonlinear system are estimated from simulated or measured trajectories. This emphasis on the empirical makes our approach broadly applicable, as the method can be applied without having to tailor anything to the particular form of the dynamics.

The approach we propose begins by constructing empirical estimates of the observability and controllability Gramians in a high (or possibly infinite) dimensional feature space. The Gramians are simultaneously

diagonalized in order to identify directions which, in the feature space, are both the most observable and the most controllable. The assumption that a nonlinear system behaves linearly when lifted to a feature space is far more reasonable than assuming linearity in the original space, and then carrying out the linear theory hoping for the best. Working in the high dimensional feature space allows one to perform linear operations on a representation of the system's state and output which can capture strong nonlinearities. Therefore a system which is not model reducible using existing methods, may become reducible when mapped into such a nonlinear feature space. This situation closely parallels the problem of linear separability in data classification: A dataset which is not linearly separable might be easily separated when mapped into a nonlinear feature space. The decision boundary is linear in this feature space, but is nonlinear in the original data space.

Nonlinear reduction of the state space already opens the door to the design of simpler controllers, but is only half of the picture. One would also like to be able to write a closed, reduced dynamical system whose input-output behavior closely captures that of the original system. This problem is the focus of the second half of our paper, where we again exploit helpful properties of RKHS in order to provide such a closed system.

The paper is organized as follows. In the next section we provide the relevant background for model reduction and balancing. We then adapt and extend balancing techniques described in the background to the current RKHS setting in Section III. Section IV then proposes a method for determining a closed, reduced nonlinear control system in light of the reduction map described in Section III. Finally, Section V provides experiments illustrating an application of the proposed methods to a specific nonlinear system.

II. BACKGROUND

Several approaches have been proposed for the reduction of linear control systems in view of control, but few exist for finite or infinite-dimensional controlled nonlinear dynamical systems. For linear systems the pioneering "Input- Output balancing" approach proposed by B.C. Moore observes that the important states are the ones that are both easy to reach and that generate a lot of energy at the output. If a large amount of energy is required to reach a certain state but the same state yields a small output energy, the state is unimportant for the input-output behavior of the system. The goal is then to find the states that are *both* the most controllable and the most observable. One way to determine such states is to find a change of coordinates where the controllability and observability Gramians (which can be viewed as a measure of the controllability and the observability of the system) are equal and diagonal. States that are difficult to reach and that don't significantly affect the output are then ignored or truncated. A system expressed in the coordinates where each state is equally controllable and observable is called its *balanced realization*.

A proposal for generalizing this approach to nonlinear control systems was advanced by J. Scherpen [24], where suitably defined controllability and observability energy functions reduce to Gramians in the linear case. In general, to find the balanced realization of a system one needs to solve a set of Hamilton-Jacobi and Lyapunov equations (as we will discuss below). Moore [19] proposed an alternative, data-based approach for balancing in the linear case. This method uses samples of the impulse response of a linear system to construct empirical controllability and observability Gramians which are then balanced and truncated using Principal Components Analysis (PCA, or POD). This data-driven strategy was then extended to nonlinear control systems with a stable linear approximation by Lall et al. [15], by effectively applying Moore's method to a nonlinear system by way of the Galerkin projection. Despite the fact that the balancing theory underpinning their approach assumes a linear system, Lall and colleagues were able to effectively reduce some nonlinear systems.

Phillips et al. [22] has also studied reduction of nonlinear circuit models in the case of linear but unbalanced coordinate transformations and found that approximation using a polynomial RKHS could afford computational advantages. Gray and Verriest mention in [8] that studying algebraically defined Gramian operators in RKHS may provide advantageous approximation properties, though the idea is not further explored. Finally, Coifman et al. [4] discuss reduction of an uncontrolled stochastic Langevin

system. There, eigenfunctions of a combinatorial Laplacian, built from samples of trajectories, provide a set of reduction coordinates but does not provide a reduced system. This method is related to kernel principal components (KPCA) using a Gaussian kernel, however reduction in this study is carried out on a simplified linear system outside the context of control.

In the following section we review balancing of linear and nonlinear systems as introduced in [19] and [24].

A. Balancing of Linear Systems

Consider a linear control system

$$\begin{aligned}\dot{x} &= Fx + Gu, \\ y &= Hx,\end{aligned}\tag{1}$$

where (F, G) is controllable, (F, H) is observable and F is Hurwitz. We define the controllability and the observability Gramians as, respectively,

$$\begin{aligned}W_c &= \int_0^\infty e^{Ft} G G^\top e^{F^\top t} dt, \\ W_o &= \int_0^\infty e^{F^\top t} H^\top H e^{Ft} dt.\end{aligned}$$

These two matrices can be viewed as a measure of the controllability and the observability of the system [19]. For instance, consider the past energy [24], $L_c(x_0)$, defined as the minimal energy required to reach x_0 from 0 in infinite time

$$L_c(x_0) = \inf_{\substack{u \in L_2(-\infty, 0), \\ x(-\infty)=0, x(0)=x_0}} \frac{1}{2} \int_{-\infty}^0 \|u(t)\|^2 dt,\tag{2}$$

and the future energy [24], $L_o(x_0)$, defined as the output energy generated by releasing the system from its initial state $x(t_0) = x_0$, and zero input $u(t) = 0$ for $t \geq 0$, i.e.

$$L_o(x_0) = \frac{1}{2} \int_0^\infty \|y(t)\|^2 dt,\tag{3}$$

for $x(t_0) = x_0$ and $u(t) = 0, t \geq 0$. In the linear case, it can be shown that $L_c(x_0) = \frac{1}{2} x_0^\top W_c^{-1} x_0$, and $L_o(x_0) = \frac{1}{2} x_0^\top W_o x_0$. The columns of W_c span the controllable subspace while the nullspace of W_o coincides with the unobservable subspace. As such, W_c and W_o (or their estimates) are the key ingredients in many model reduction techniques. It is also well known that W_c and W_o satisfy the Lyapunov equations [19]

$$\begin{aligned}FW_c + W_c F^\top &= -GG^\top, \\ F^\top W_o + W_o F &= -H^\top H.\end{aligned}$$

Several methods have been developed to solve these equations directly [16], [17].

The idea behind balancing is to find a representation where the system's observable and controllable subspaces are aligned so that reduction, if possible, consists of eliminating uncontrollable states which are also the least observable. More formally, we would like to find a new coordinate system such that

$$W_c = W_o = \Sigma = \text{diag}\{\sigma_1, \dots, \sigma_n\},$$

where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$. If (F, G) is controllable and (F, H) is observable, then there exists a transformation such that the state space expressed in the transformed coordinates (TFT^{-1}, TG, HT^{-1}) is balanced and $TW_c T^\top = T^{-\top} W_o T^{-1} = \Sigma$. Typically one looks for a gap in the singular values $\{\sigma_i\}$ for guidance as to where truncation should occur. If we see that there is a k such that $\sigma_k \gg \sigma_{k+1}$, then the states most responsible for governing the input-output relationship of the system are (x_1, \dots, x_k) while (x_{k+1}, \dots, x_n) are assumed to make negligible contributions.

If A is unstable then the controllability and observability quantities defined in (2) are undefined since the integrals will be unbounded. There may, however, still exist solutions to the Lyapunov equations (4)

when A is unstable, and these solutions will be unique if and only if $\lambda(A) + \lambda(\bar{A}) \neq 0$. In this case balancing may be carried out as usual by finding (if possible) a transformation T such that $W_c = W_o = \Sigma$ where Σ is again diagonal and positive semidefinite [26], [10]. Other approaches to balancing unstable linear systems exist (see [9], [28] for the method of LQG balancing for example).

Although several methods also exist for computing T [16], [17], it is common to simply compute the Cholesky decomposition of W_o so that $W_o = ZZ^\top$, and form the SVD $U\Sigma^2U^\top$ of $Z^\top W_c Z$. Then T is given by $T = \Sigma^{\frac{1}{2}}U^\top Z^{-1}$. We also note that the problem of finding the coordinate change T can be seen as an optimization problem [1] of the form

$$\min_T \text{trace}[TW_cT^* + T^{-*}W_oT^{-1}].$$

B. Balancing of Nonlinear Systems

In the nonlinear case, the energy functions L_c and L_o in (2) and (3) are obtained by solving both a Lyapunov and a Hamilton-Jacobi equation. Here we follow the development of Scherpen [24]. Consider the nonlinear system

$$\begin{cases} \dot{x} &= f(x) + \sum_{i=1}^m g_i(x)u_i, \\ y &= h(x), \end{cases} \quad (4)$$

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, $f(0) = 0$, $g_i(0) = 0$ for $1 \leq i \leq m$, and $h(0) = 0$. Moreover, assume the following Hypothesis.

Assumption A: The linearization of (4) around the origin is controllable, observable and $F = \frac{\partial f}{\partial x}|_{x=0}$ is asymptotically stable.

Theorem 2.1: [24] If the origin is an asymptotically stable equilibrium of $f(x)$ on a neighborhood W of the origin, then for all $x \in W$, $L_o(x)$ is the unique smooth solution of

$$\frac{\partial L_o}{\partial x}(x)f(x) + \frac{1}{2}h^\top(x)h(x) = 0, \quad L_o(0) = 0 \quad (5)$$

under the assumption that (5) has a smooth solution on W . Furthermore for all $x \in W$, $L_c(x)$ is the unique smooth solution of

$$\frac{\partial L_c}{\partial x}(x)f(x) + \frac{1}{2}\frac{\partial L_c}{\partial x}(x)g(x)g^\top(x)\frac{\partial^\top L_c}{\partial x}(x) = 0, \quad L_c(0) = 0 \quad (6)$$

under the assumption that (6) has a smooth solution \bar{L}_c on W and that the origin is an asymptotically stable equilibrium of $-(f(x) + g(x)g^\top(x)\frac{\partial \bar{L}_c}{\partial x}(x))$ on W .

With the controllability and the observability functions on hand, the input-normal/output-diagonal realization of system (4) can be computed by way of a coordinate transformation. More precisely,

Theorem 2.2: [24] Consider system (4) under Assumption A and the assumptions in Theorem 2.1. Then, there exists a neighborhood W of the origin and coordinate transformation $x = \varphi(z)$ on W converting the energy functions into the form

$$\begin{aligned} L_c(\varphi(z)) &= \frac{1}{2}z^\top z, \\ L_o(\varphi(z)) &= \frac{1}{2}\sum_{i=1}^n z_i^2 \sigma_i(z_i)^2, \end{aligned}$$

where $\sigma_1(x) \geq \sigma_2(x) \geq \dots \geq \sigma_n(x)$. The functions $\sigma_i(\cdot)$ are called *Hankel singular value functions*. Analogous to the linear case, the system's states can be sorted in order of importance by sorting the singular value functions, and reduction proceeds by removing the least important states.

In the above framework for balancing of nonlinear systems, one needs to solve (or numerically evaluate) the PDEs (5), (6) and compute the coordinate change $x = \varphi(z)$, however there are no systematic methods or tools for solving these problems. Various approximate solutions based on Taylor series expansions have

been proposed [13], [12], [7]. Newman and Krishnaprasad [20] introduce a statistical approximation based on exciting the system with white Gaussian noise and then computing the balancing transformation using an algorithm from differential topology. As mentioned earlier, an essentially linear empirical approach was proposed in [15]. In this paper, we combine aspects of both data-driven approaches and analytic approaches by carrying out balancing in a suitable RKHS.

III. EMPIRICAL BALANCING OF NONLINEAR SYSTEMS IN RKHS

We consider a general nonlinear system of the form

$$\begin{cases} \dot{x} &= f(x, u) \\ y &= h(x) \end{cases} \quad (7)$$

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, $f(0, 0) = 0$, and $h(0) = 0$. Let $\mathcal{R}(x_0) = \{x' \in \mathbb{R}^n : \exists u \in L_\infty(\mathbb{R}, \mathbb{R}^m) \text{ and } \exists T \in [0, \infty) \text{ such that } x(0) = x_0 \text{ and } x(T) = x'\}$ be the reachable set from the initial condition $x(0) = x_0$.

Hypothesis H: The system (7) is zero-state observable, its linearization around the origin is controllable, and the origin of $\dot{x} = f(x, 0)$ is asymptotically stable.

We treat the problem of estimating the observability and controllability Gramians as one of estimating an integral operator from data in a reproducing kernel Hilbert space (RKHS) [2]. *Our approach hinges on the key modeling assumption that the nonlinear dynamical system is linear in an appropriate high (or possibly infinite) dimensional lifted feature space.* Covariance operators in this feature space and their empirical estimates are the objects of primary importance and contain the information needed to perform model reduction. In particular, the (linear) observability and controllability Gramians are estimated and diagonalized in the feature space, but capture nonlinearities in the original state space. The reduction approach we propose adapts ideas from kernel PCA (KPCA) [25] and is driven by a set of simulated or sampled system trajectories, extending and generalizing the work of Moore [19] and Lall et al. [15].

In the development below we lift state vectors of the system (7) into a reproducing kernel Hilbert space [2].

A. Elements of Learning Theory

In this section, we give a brief overview of reproducing kernel Hilbert spaces as used in statistical learning theory. The discussion here borrows heavily from [5], [27]. Early work developing the theory of RKHS was undertaken by N. Aronszajn [2].

Definition 3.1: Let \mathcal{H} be a Hilbert space of functions on a set \mathcal{X} . Denote by $\langle f, g \rangle$ the inner product on \mathcal{H} and let $\|f\| = \langle f, f \rangle^{1/2}$ be the norm in \mathcal{H} , for f and $g \in \mathcal{H}$. We say that \mathcal{H} is a reproducing kernel Hilbert space (RKHS) if there exists $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that

- i. K has the reproducing property, i.e. $\forall f \in \mathcal{H}, f(x) = \langle f(\cdot), K(\cdot, x) \rangle$.
- ii. K spans \mathcal{H} , i.e. $\mathcal{H} = \text{span}\{K(x, \cdot) | x \in \mathcal{X}\}$.

K will be called a reproducing kernel of \mathcal{H} . $\mathcal{H}_K(X)$ will denote the RKHS \mathcal{H} with reproducing kernel K .

Definition 3.2: (Mercer kernel map) A function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a Mercer kernel if it is continuous, symmetric and positive definite.

The important properties of reproducing kernels are summarized in the following proposition

Proposition 3.1: If K is a reproducing kernel of a Hilbert space \mathcal{H} , then

- i. $K(x, y)$ is unique.
- ii. $\forall x, y \in \mathcal{X}, K(x, y) = K(y, x)$ (symmetry).
- iii. $\sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j) \geq 0$ for $\alpha_i \in \mathbb{R}$ and $x_i \in \mathcal{X}$ (positive definiteness).
- iv. $\langle K(x, \cdot), K(y, \cdot) \rangle_{\mathcal{H}} = K(x, y)$.

- v. Let $c \neq 0$. The following kernels, defined on a compact domain $\mathcal{X} \subset \mathbb{R}^n$, are Mercer kernels:
 $K(x, y) = x \cdot y'$ (Linear), $K(x, y) = (1 + x \cdot y)^d$, $d \in \mathbb{N}$ (Polynomial), $K(x, y) = e^{-\frac{\|x-y\|^2}{\sigma^2}}$, $\sigma > 0$ (Gaussian).

Theorem 3.1: Let $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a symmetric and positive definite function. Then, there exists a Hilbert space of functions \mathcal{H} defined on \mathcal{X} admitting K as a reproducing Kernel.

Conversely, let \mathcal{H} be a Hilbert space of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ satisfying

$$\forall x \in \mathcal{X}, \exists \kappa_x > 0, \quad \text{such that} \quad |f(x)| \leq \kappa_x \|f\|_{\mathcal{H}}, \quad \forall f \in \mathcal{H}.$$

Then, \mathcal{H} has a reproducing kernel K .

Theorem 3.2: Every sequence of functions $(f_n)_{n \geq 1}$ which converges strongly to a function f in $\mathcal{H}_K(X)$, converges also in the pointwise sense, that is, $\lim_{n \rightarrow \infty} f_n(x) = f(x)$, for any point $x \in X$. Further, this convergence is uniform on every subset of X on which $x \mapsto K(x, x)$ is bounded.

Theorem 3.3: Let $K(x, y)$ be a positive definite kernel on a compact domain or a manifold X . Then there exists a Hilbert space \mathcal{F} and a function $\Phi : X \rightarrow \mathcal{F}$ such that

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle \quad \text{for } x, y \in X.$$

Φ is called a feature map, and \mathcal{F} a feature space¹.

Remarks.

- i. In theorem 3.3, and using property [iv.] in Proposition 3.1, we can take $\Phi(x) := K_x := K(x, \cdot)$ in which case $\mathcal{F} = \mathcal{H}$ – the “feature space” is the RKHS.
- ii. The fact that Mercer kernels are positive definite and symmetric reminds us of similar properties of Gramians and covariance matrices. This is an essential fact that we are going to use in the following.
- iii. In practice, we choose a Mercer kernel, such as the ones in [v.] in Proposition 3.1, and theorem 3.1 guarantees the existence of a Hilbert space admitting such a function as a reproducing kernel.

◁

RKHS play an important role in learning theory whose objective is to find an unknown function $f : X \rightarrow Y$ from random samples $(x_i, y_i)_{i=1}^m$. For instance, assume that the random probability measure that governs the random samples is ρ and is defined on $Z := X \times Y$. Let X be a compact subset of \mathbb{R}^n and $Y = \mathbb{R}$. If we define the least square error of f as

$$\mathcal{E} = \int_{X \times Y} (f(x) - y)^2 d\rho, \quad (8)$$

then the function that minimizes the error is the regression function f_ρ

$$f_\rho(x) = \int_{\mathbb{R}} y d\rho(y|x), \quad x \in X,$$

where $\rho(y|x)$ is the conditional probability measure on \mathbb{R} . Since ρ is unknown, neither f_ρ nor \mathcal{E} is computable. We only have the samples $s := (x_i, y_i)_{i=1}^m$. The error f_ρ is approximated by the empirical error $\mathcal{E}_s(f)$ by

$$\mathcal{E}_s(f) = \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2, \quad (9)$$

for $\lambda \geq 0$, λ plays the role of a regularizing parameter. In learning theory, the minimization is taken over functions from a hypothesis space often taken to be a ball of a RKHS \mathcal{H}_K associated to Mercer kernel K , and the function f_s that minimizes the empirical error \mathcal{E}_s is

$$f_s = \sum_{j=1}^m c_j K(x, x_j), \quad (10)$$

¹The dimension of the feature space can be infinite and corresponds to the dimension of the eigenspace of the integral operator $L_K : \mathcal{L}_\nu^2(X) \rightarrow \mathcal{C}(X)$ defined as $(L_K f)(x) = \int K(x, t) f(t) d\nu(t)$ if K is a Mercer kernel, for $f \in \mathcal{L}_\nu^2(X)$ and ν is a Borel measure on X .

where the coefficients $(c_j)_{j=1}^m$ is solved by the linear system

$$\lambda m c_i + \sum_{j=1}^m K(x_i, x_j) c_j = y_i, \quad i = 1, \dots, m, \quad (11)$$

and f_s is taken as an approximation of the regression function f_ρ . Hence, minimizing over the (possibly infinite dimensional) Hilbert space, reduces to minimizing over \mathbb{R}^m . The series (10) converges absolutely and uniformly to f .

We call *learning* the process of approximating the unknown function f from random samples on Z .

In the following, we assume that the kernels K are continuous and bounded by

$$\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x, x)} < \infty.$$

B. Empirical Gramians in RKHS

Following [19], we estimate the controllability Gramian by exciting each coordinate of the input with impulses while setting $x_0 = 0$. One can also further excite using rotations of impulses as suggested in [15], however for simplicity we consider only the original signals proposed in [19]. Let $u^i(t) = \delta(t)e_i$ be the i -th excitation signal, and let $x^i(t)$ be the corresponding response of the system. Form the matrix $X(t) = [x^1(t) \cdots x^m(t)] \in \mathbb{R}^{n \times m}$, so that $X(t)$ is seen as a data matrix with column observations given by the respective responses $x^i(t)$. Then $W_c \in \mathbb{R}^{n \times n}$ is given by

$$W_c = \frac{1}{m} \int_0^\infty X(t) X(t)^\top dt. \quad (12)$$

We can approximate this integral by sampling the matrix function $X(t)$ within a finite time interval $[0, T]$ assuming the regular partition $\{t_i\}_{i=1}^N, t_i = (T/N)i$. This leads to the empirical controllability Gramian

$$\widehat{W}_c = \frac{T}{mN} \sum_{i=1}^N X(t_i) X(t_i)^\top. \quad (13)$$

As described in [19], the observability Gramian is estimated by fixing $u(t) = 0$, setting $x_0 = e_i$ for $i = 1, \dots, n$, and measuring the corresponding system output responses $y^i(t)$. As before, assemble the responses into a matrix $Y(t) = [y^1(t) \cdots y^n(t)] \in \mathbb{R}^{p \times n}$. The observability Gramian $W_o \in \mathbb{R}^{n \times n}$ and its empirical counterpart \widehat{W}_o are given by

$$W_o = \frac{1}{p} \int_0^\infty Y(t)^\top Y(t) dt \quad (14)$$

and

$$\widehat{W}_o = \frac{T}{pN} \sum_{i=1}^N \tilde{Y}(t_i) \tilde{Y}(t_i)^\top \quad (15)$$

where $\tilde{Y}(t) = Y(t)^\top$. The matrix $\tilde{Y}(t_i) \in \mathbb{R}^{n \times p}$ can be thought of as a data matrix with column observations

$$d_j(t_i) = (y_j^1(t_i), \dots, y_j^n(t_i))^\top \in \mathbb{R}^n, \quad j = 1, \dots, p, \quad i = 1, \dots, N \quad (16)$$

so that $d_j(t_i)$ corresponds to the response at time t_i of the single output coordinate j to each of the (separate) initial conditions $x_0 = e_k, k = 1, \dots, n$. This convention will lead to greater clarity in the steps that follow.

C. Kernel PCA

Kernel PCA [25] will be a helpful starting point for understanding the approach to balanced reduction introduced in this paper. We briefly review the relevant background here. Kernel PCA (KPCA) generalizes linear PCA by carrying out PCA in a high dimensional feature space defined by a feature map $\Phi : \mathbb{R}^n \rightarrow \mathcal{F}$. Taking the feature map $\Phi(x) = K_x$ and given the set of data $\mathbf{x} := \{x_i\}_{i=1}^N \in \mathbb{R}^n$, we can consider PCA in the feature space by simply working with the covariance of the mapped vectors,

$$C_{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \Phi(x_i) \otimes \Phi(x_i), \quad (17)$$

where $\Phi(x_i) \otimes \Phi(x_i) = \langle \Phi(x_i), \cdot \rangle \Phi(x_i)$ denotes the tensor product between two vectors in \mathcal{H} . We will assume the data are centered in the feature space so that $\sum_i \Phi(x_i) = 0$. If not, data may be centered according to the prescription in [25].

The principal subspaces are computed by diagonalizing $C_{\mathbf{x}}$, however as is shown in [25], one can equivalently form the matrix $K \in \mathbb{R}^{N \times N}$ of kernel products $(K)_{ij} = K(x_i, x_j)$ for $i, j = 1, \dots, N$, and solve the eigenproblem

$$K\alpha = N\lambda\alpha. \quad (18)$$

If

$$C_{\mathbf{x}}v_i = \lambda_i v_i, \quad (19)$$

then we have that

$$v_i = \Psi\alpha_i \quad (20)$$

where $\Psi := (\Phi(x_1) \cdots \Phi(x_N))$, and the non-zero eigenvalues of K and $C_{\mathbf{x}}$ coincide.

The eigenvectors α_i of K are then normalized so that the eigenvectors v_i of $C_{\mathbf{x}}$ have unit norm in the feature space, leading to the condition $\|\alpha_i\|^2 = \lambda_i^{-1}$. Assuming this normalization convention, sort the eigenvectors according to the magnitudes of the corresponding eigenvalues in descending order, and form the matrix

$$A_q = [\alpha_1 \cdots \alpha_q], 1 \leq q \leq \min(n, N). \quad (21)$$

Similarly, form the matrix $V_q = [v_1 \cdots v_q], 1 \leq q \leq n$ of sorted eigenvectors of $C_{\mathbf{x}}$. The first q principal components of a vector $x = \Phi(\tilde{x})$ in the feature space are then given by $V_q^\top x$. It can be shown however (see [25]) that principal components in the feature space can be computed in the original space with kernels using the map

$$\Pi(x) := A_q^\top \mathbf{k}(x), \quad (22)$$

where $\mathbf{k}(x) = (K(x, x_1), \dots, K(x, x_N))^\top$.

D. Model Order Reduction Map

The method we propose consists, in essence, of collecting samples and then performing a process similar to “simultaneous principal components analysis” on the controllability and observability Gramian estimates in the (same) RKHS. As mentioned above, given a choice of the kernel K defining a RKHS \mathcal{H} , principal components in the feature space can be computed implicitly in the original input space using K . It is worth emphasizing however that we will be co-diagonalizing *two* Gramians in the feature space by way of a *non-orthogonal* transformation; the process bears a resemblance to (K)PCA, and yet is distinct. Indeed the favorable properties associated with an orthonormal basis are no longer available, the quantities we will in practice diagonalize are different, and the issue of data-centering must be considered with some additional care.

First note that the controllability Gramian \widehat{W}_c can be viewed as the sample covariance of a collection of $N \cdot m$ vectors, scaled by T

$$\widehat{W}_c = \frac{T}{mN} \sum_{i=1}^N X(t_i) X(t_i)^\top = \frac{T}{mN} \sum_{i=1}^N \sum_{j=1}^m x^j(t_i) x^j(t_i)^\top \quad (23)$$

and the observability Gramian can be similarly viewed as the sample covariance of a collection of $N \cdot p$ vectors

$$\widehat{W}_o = \frac{T}{pN} \sum_{i=1}^N \sum_{j=1}^p d_j(t_i) d_j(t_i)^\top \quad (24)$$

where the d_j are defined in Equation (16).

We can thus consider three quantities of interest:

- The *controllability kernel matrix* $K_c \in \mathbb{R}^{Nm \times Nm}$ of kernel products

$$(K_c)_{\mu\nu} = K(x_\mu, x_\nu) = \langle \Phi(x_\mu), \Phi(x_\nu) \rangle_{\mathcal{F}} \quad (25)$$

for $\mu, \nu = 1, \dots, Nm$ where we have re-indexed the set of vectors $\{x^j(t_i)\}_{i,j} = \{x_\mu\}_\mu$ to use a single linear index.

- The *observability kernel matrix* $K_o \in \mathbb{R}^{Np \times Np}$,

$$(K_o)_{\mu\nu} = K(d_\mu, d_\nu) = \langle \Phi(d_\mu), \Phi(d_\nu) \rangle_{\mathcal{F}} \quad (26)$$

for $\mu, \nu = 1, \dots, Np$, where we have again re-indexed the set $\{d_j(t_i)\}_{i,j} = \{d_\mu\}_\mu$ for simplicity.

- The *Hankel kernel matrix* $K_{o,c} \in \mathbb{R}^{Np \times Nm}$,

$$(K_{o,c})_{\mu\nu} = K(d_\mu, x_\nu) = \langle \Phi(d_\mu), \Phi(x_\nu) \rangle_{\mathcal{F}} \quad (27)$$

for $\mu = 1, \dots, Np$, $\nu = 1, \dots, Nm$.

We have chosen the suggestive terminology ‘‘Hankel kernel matrix’’ above because the square-roots of the nonzero eigenvalues of the matrix $K_{o,c} K_{o,c}^\top$ are the empirical Hankel singular values of the system mapped into feature space, where we assume the system behaves linearly. This assertion will be proved immediately below. Note that ordinarily, $Nm, Np \gg n$ and K_c, K_o will be rank deficient.

Before proceeding we consider the issue of data centering in feature space. PCA and kernel PCA assume that the data have been centered in order to make the problem translation invariant. In the setting considered here, we have two distinct sets of data: the observability samples and the controllability samples. A reasonable centering convention centers the data in each of these datasets separately. Let Ψ denote the matrix whose columns are the observability samples mapped into feature space by Φ , and let Φ be the matrix similarly built from the feature space representation of the controllability samples. Then $K_o = \Psi^\top \Psi$, $K_c = \Phi^\top \Phi$ and $K_{o,c} = \Psi^\top \Phi$. Assume for the moment that there are M observability data samples and N controllability samples, and let $\mathbf{1}_N, \mathbf{1}_M$ denote the length N, M vectors of all ones, respectively. We can define centered versions of the feature space data matrices Φ, Ψ as

$$\tilde{\Phi} = \Phi - \mu_c \mathbf{1}_N^\top, \quad \tilde{\Psi} = \Psi - \mu_o \mathbf{1}_M^\top \quad (28)$$

where $\mu_c := N^{-1} \Phi \mathbf{1}_N$ and $\mu_o := M^{-1} \Psi \mathbf{1}_M$. We will need two centered quantities in the development below. The first centered quantity we consider is the centered version of $K_{o,c}$, namely $\tilde{K}_{o,c} = \tilde{\Psi}^\top \tilde{\Phi}$. Although one cannot compute μ_c, μ_o explicitly from the data, we can compute $\tilde{K}_{o,c}$ by observing that

$$\begin{aligned} \tilde{K}_{o,c} &= (\Psi - \mu_o \mathbf{1}_M^\top)^\top (\Phi - \mu_c \mathbf{1}_N^\top) \\ &= K_{o,c} - \frac{1}{N} K_{o,c} \mathbf{1}_N \mathbf{1}_N^\top - \frac{1}{M} \mathbf{1}_M \mathbf{1}_M^\top K_{o,c} + \frac{1}{NM} \mathbf{1}_M \mathbf{1}_M^\top K_{o,c} \mathbf{1}_N \mathbf{1}_N^\top. \end{aligned} \quad (29)$$

The second quantity we’ll need a centered version of the *empirical observability feature map*

$$\mathbf{k}_o(x) := \Psi^\top \Phi(x) = (K(x, d_1), \dots, K(x, d_M))^\top \quad (30)$$

where $x \in \mathbb{R}^d$ is the state variable and the observability samples $\{d_j\}$ are again indexed by a single variable as in Equation (26). Centering follows reasoning similar to that of the Hankel kernel matrix immediately above:

$$\begin{aligned}\tilde{\mathbf{k}}_o(x) &= (\Psi - \mu_o \mathbf{1}_M^\top)^\top (\Phi(x) - \mu_c) \\ &= \mathbf{k}_o(x) - \frac{1}{N} K_{o,c} \mathbf{1}_N - \frac{1}{M} \mathbf{1}_M \mathbf{1}_M^\top \mathbf{k}_o(x) + \frac{1}{NM} \mathbf{1}_M \mathbf{1}_M^\top K_{o,c} \mathbf{1}_N.\end{aligned}\quad (31)$$

Note: Throughout the remainder of this paper we will drop the special notation $\tilde{K}_{o,c}, \tilde{\mathbf{k}}_o(x)$ and assume that $K_{o,c}, \mathbf{k}_o(x)$ are centered appropriately.

With the quantities defined above, we can co-diagonalize the empirical Gramians (balancing) and reduce the dimensionality of the state variable (truncation) in feature space by carrying out calculations in the original data space. As the system is assumed to behave linearly in the feature space, the order of the model can be reduced by discarding small Hankel values $\{\Sigma_{ii}\}_{i=q+1}^n$, and projecting onto the subspace associated with the first $q < n$ largest eigenvalues. The following key result describes this process:

Theorem 3.4 (Balanced Reduction in Feature Space): Balanced reduction in feature space can be accomplished by applying the state-space reduction map $\Pi : \mathbb{R}^n \rightarrow \mathbb{R}^q$ given by

$$\Pi(x) = T_q^\top \mathbf{k}_o(x), \quad x \in \mathbb{R}^n \quad (32)$$

where $T_q = V_q \Sigma_q^{-1/2}$ if $K_{o,c} K_{o,c}^\top = V \Sigma^2 V^\top$, and $\mathbf{k}_o(x)$ is the empirical observability feature map.

Proof: We assume the data have been centered in feature space. Let Φ be a matrix with columns $\{\Phi(x^j(x_i))\}, i = 1, \dots, N, j = 1, \dots, m$, so that $X = \Phi \Phi^\top$ is the feature space controllability Gramian counterpart to Equation (23). Similarly, let Ψ be a matrix with columns $\{\Phi(d_j(t_i))\}, i = 1, \dots, N, j = 1, \dots, p$, so that $Y = \Psi \Psi^\top$ is the feature space observability Gramian counterpart to Equation (24). Since by definition $K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{F}}$, we also have that $K_c = \Phi^\top \Phi$ and $K_o = \Psi^\top \Psi$. In general the Gramians X, Y are infinite dimensional whereas the kernel matrices K_c, K_o are necessarily of finite dimension.

We now carry out linear balancing on (X, Y) in the feature space (RKHS). First, take the SVD of $X^{1/2} \Psi$ so that

$$U \Sigma V^\top = X^{1/2} \Psi \quad (33)$$

$$U \Sigma^2 U^\top = (X^{1/2} \Psi)(X^{1/2} \Psi)^\top = X^{1/2} Y X^{1/2} \quad (34)$$

$$V \Sigma^2 V^\top = (X^{1/2} \Psi)^\top (X^{1/2} \Psi) = \Psi^\top \Phi \Phi^\top \Psi = K_{o,c} K_{o,c}^\top. \quad (35)$$

The last equality in Equation (34) follows since X is symmetric and therefore $X^{1/2}$ is too. The linear balancing transformation is then given by $M = \Sigma^{1/2} U^\top X^{-1/2}$, and one can readily verify that $M X M^\top = M^{-\top} Y M^{-1} = \Sigma$. Here, inverses should be interpreted as pseudo-inverses when appropriate². From Equations (33)-(35), we see that $U^\top = \Sigma^{-1} V^\top \Psi^\top X^{1/2}$ and thus $M = \Sigma^{-1/2} V^\top \Psi^\top$. We can project an arbitrary mapped data point $\Phi(x)$ onto the (balanced) ‘‘principal’’ subspace of dimension q spanned by the first q rows of M by computing

$$M_q \Phi(x) = \Sigma_q^{-1/2} V_q^\top \Psi^\top \Phi(x) = \Sigma_q^{-1/2} V_q^\top \mathbf{k}_o(x) \quad (36)$$

where $\mathbf{k}_o(x) := \Psi^\top \Phi(x)$ is the empirical observability feature map, recalling that V_q is the matrix formed by taking the top q eigenvectors of $K_{o,c} K_{o,c}^\top$ by Equation (35). ■

We note that square roots of the non-zero eigenvalues of $K_{o,c} K_{o,c}^\top$ are exactly the Hankel singular values of the system mapped into the feature space, under the assumption of linearity in the feature space. This can be seen by noting that $\lambda_+(YX) = \lambda_+(X^{1/2} Y X^{1/2}) = \lambda_+(K_{o,c} K_{o,c}^\top)$, where $\lambda_+(\cdot)$ refers to the non-zero eigenvalues of its argument.

In Section IV below we show how to use the nonlinear reduction map (32) to realize a closed, reduced order system which can approximate the original system to a high degree of accuracy.

²Such as in the case of $X^{-1/2}$ when the number of data points is less than the dimension of the RKHS.

IV. CLOSED DYNAMICS OF THE REDUCED SYSTEM

Given the nonlinear state space reduction map $\Pi : \mathbb{R}^n \rightarrow \mathbb{R}^q$, a remaining challenge is to construct a corresponding (reduced) dynamical system on the reduced state space which well approximates the input-output behavior of the original system on the original state space. Setting $x_r = \Pi(x)$ and applying the chain rule,

$$\dot{x}_r = \left(J_\Pi(x) f(x, u) \right) \Big|_{x=\Pi^\dagger(x_r)} \quad (37)$$

where Π^\dagger refers to an appropriate notion (to be defined) of the inverse of Π . However we are faced with the difficulty that the map Π is not in general injective (even if $q = n$), and moreover one cannot guarantee that an arbitrary point in the RKHS has a non-empty preimage under Φ [18]. We propose an approximation scheme to get around this difficulty: The dynamics f will be approximated by an element of an RKHS *defined on the reduced state space*. When f is assumed to be known explicitly it can be approximated to a high degree of accuracy. An approximate, least-squares notion of Π^\dagger will be given to first or second order via a Taylor series expansion, but only where it is strictly needed – and at the last possible moment – so that a first or second order approximation will not be as crude as one might suppose. We will also consider, as an alternative, a direct approximation of $J_\Pi(\Pi^\dagger(x_r))$ which takes into account further properties of the reproducing kernel as well as the fact that the Jacobian is to be evaluated at $x = \Pi^\dagger(x_r)$ in particular. In both cases, the important ability of the map Π to capture strong nonlinearities will not be significantly diminished.

A. Representation of the dynamics in RKHS

The vector-valued map $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ can be approximated by a composing a set of n regression functions (one for each coordinate) $\hat{f}_i : \mathbb{R}^{q \times m} \rightarrow \mathbb{R}$ in an RKHS, with the reduction map Π . It is reasonable to expect that this approximation will be better than directly computing $f(\Pi^\dagger(x_r), u)$ using, for instance, a Taylor expansion approximation for Π^\dagger which may ignore important nonlinearities at a stage where crude approximations must be avoided.

Let $\tilde{x} = \Pi(x)$ denote a reduced state variable, and concatenate the input examples $\tilde{x}_j = \Pi(x_j) \in \mathbb{R}^q, u_j \in \mathbb{R}^m$ so that $z_j = (\tilde{x}_j, u_j) \in \mathbb{R}^{q \times m}$, and $\{(f_i(x_j, u_j), z_j)\}_{j=1}^\ell$ is a set of input-output training pairs describing the i -th coordinate of the map $(\tilde{x}, u) \mapsto f(x, u)$. The training examples should characterize “typical” behaviors of the system, and can even re-use those trajectories simulated in response to impulses for estimating the Gramians above. We will seek the function $\hat{f}_i \in \mathcal{H}$ which minimizes

$$\sum_{j=1}^{\ell} (\hat{f}_i(z_j) - f_i(x_j, u_j))^2 + \lambda_i \|\hat{f}_i\|_{\mathcal{H}}^2$$

where λ_i here is a regularization parameter. We have chosen the square loss, however other suitable loss functions may be used. It can be shown [27] that in this case \hat{f}_i takes the form $\hat{f}_i(z) = \sum_{j=1}^{\ell} c_j^i K^f(z, z_j)$, $i = 1, \dots, n$, where K^f defines the RKHS \mathcal{H}_f (and is unrelated to K used to estimate the Gramians). Note that although our notation takes the RKHS for each coordinate function to be the same, in general this need not be true: different kernels may be chosen for each function. Here the $\{c_j^i\}$ comprise a set of coefficients learned using the regularized least squares (RLS) algorithm. The kernel family and any hyper-parameters can be chosen by cross-validation. For notational convenience we will further define the vector-valued empirical feature map

$$(\mathbf{k}^f(\tilde{x}, u))_i := K^f((\tilde{x}, u), z_i)$$

for $i = 1, \dots, \ell$. In this notation $\hat{f}_i(\Pi(x), u) = \mathbf{c}_i^\top \mathbf{k}^f(\tilde{x}, u)$ where $(\mathbf{c}_i)_j = c_j^i$.

A broad class of systems seen in the literature [24] are also characterized by separable dynamics of the form $\dot{x} = f(x) + \sum_{i=1}^m g_i(x) u_i$. In this case one need only estimate the functions f and g_i from examples $\{(\Pi(x_j), f(x_j))\}_j$ and $\{(\Pi(x_j), g(x_j))\}_j$.

B. Approximation of the Jacobian Contribution

We turn to approximating the component $J_\Pi(\Pi^\dagger(x_r))$ appearing in Equation (37).

1) *Inverse-Taylor Expansion:* A simple solution is to compute a low-order Taylor expansion of Π and then invert it using the Moore-Penrose pseudoinverse to obtain the approximation. For example, consider the first order expansion $\Pi(x) \approx \Pi(a) + J_\Pi(a)(x - a)$. Then we can approximate $\Pi^\dagger(x_r)$ (in the first-order, least-norm sense) as

$$\hat{\Pi}^\dagger(x_r) := (J_\Pi(a))^\dagger(x_r - \Pi(a)) + a. \quad (38)$$

We may start with $a = x_0$, but periodically update the expansion in different regions of the dynamics if desired. A good expansion point could be the estimated preimage of $x_r(t)$ returned by the algorithm proposed in [14]. If $\mathbf{k}_o(x)$ is the centered version of the length M vector $\mathbf{k}_o(x)$ defined by (30), then

$$J_\Pi = \frac{\partial \Pi}{\partial x} = T_q^\top \left(I - \frac{1}{M} \mathbf{1}_M \mathbf{1}_M^\top \right) \frac{\partial \mathbf{k}_o(x)}{\partial x}$$

where $\mathbf{1}_M$ is the length M vector of all ones. An example calculation of $(\partial_x \mathbf{k}_o(x))_i = \partial_x K(x, d_i)$ in the case of a polynomial kernel is given in the section immediately below.

2) *Exploiting Kernel Properties:* For certain choices of the kernel K defining the Gramian feature space \mathcal{H} , one can exploit the fact that K_x and its derivative bear a special relationship, and potentially improve the estimate for $J_\Pi(\Pi^\dagger(x_r))$. Perhaps the most commonly used off-the-shelf kernel families are the polynomial and Gaussian families. For any two kernels with hyperparameters p and q (respectively) in one of these classes, we have that $K_p = (K_q)^{p/q}$. We'll consider the polynomial kernel of degree d , $K_d(x, y) := (1 + \langle x, y \rangle)^d$ in particular; the Gaussian case can be derived using similar reasoning. For a polynomial kernel we have that

$$\frac{\partial K_d(x, y)}{\partial x} = d K_{d-1}(x, y) y^\top = d (K_d(x, y))^{\frac{d-1}{d}} y^\top.$$

Recalling that $K_d(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$ and $x_r = \Pi(x) = V_q^\top \Phi(x)$, if Π were invertible then we would have

$$\left. \frac{\partial K_d(x, y)}{\partial x} \right|_{x=\Pi^{-1}(x_r)} = d \langle (\Phi \circ \Pi^{-1})(x_r), \Phi(y) \rangle^{\frac{d-1}{d}} y^\top.$$

The map Π is not injective however, and in addition the fibers of Φ may be potentially empty, so we must settle for an approximation. It is reasonable then to *define* $(\Phi \circ \Pi^\dagger)(x_r)$ as the solution to the convex optimization problem

$$\begin{aligned} \min_{z \in \mathcal{H}} \quad & \|z\|_{\mathcal{H}} \\ \text{subj. to} \quad & \|M_q z - x_r\|_{\mathbb{R}^k} = 0 \end{aligned} \quad (39)$$

where $M_q : \mathcal{H} \rightarrow \mathbb{R}^k$ is defined as in Equation (36). If a point $z \in \mathcal{H}$ has a pre-image in \mathbb{R}^n this definition is consistent with composing Φ with the formal definition $\Phi^{-1}(z) = \{x \in \mathbb{R}^n \mid \Phi(x) = z\}$ and noting that in this case $\Pi \circ \Phi^{-1} = M_q(\Phi \circ \Phi^{-1}) = M_q z$. Furthermore, a trajectory $x_r(t)$ of the closed dynamical system on the reduced statespace need not (and may not) have a counterpart in the original statespace by virtue of the way in which Π^\dagger is used in our formulation of the reduction map and corresponding reduced dynamical system.

One will recognize that the solution z^* to (39) is just the Moore-Penrose pseudoinverse $z^* = M_q^\dagger x_r$. Inserting this solution into the feature map representation of a kernel K gives the following definition for

$K(\Pi^\dagger(x_r), y)$:

$$\begin{aligned}
K(\Pi^\dagger(x_r), y) &= \langle (\Phi \circ \Pi^\dagger)(x_r), \Phi(y) \rangle_{\mathcal{H}} \\
&= \langle M_q^\dagger x_r, \Phi(y) \rangle_{\mathcal{H}} = \langle x_r, (M_q^\top)^\dagger \Phi(y) \rangle_{\mathbb{R}^k} \\
&= \langle x_r, (M_q M_q^\top)^{-1} M_q \Phi(y) \rangle \\
&= \langle x_r, (M_q M_q^\top)^{-1} \Pi(y) \rangle \\
&= \langle x_r, (T_q^\top K_o T_q)^{-1} \Pi(y) \rangle
\end{aligned}$$

where the final equality follows applying Equations (33)-(35) and T_q is defined as in Theorem 3.4. Substituting into the derivative for a polynomial kernel $K = K_d$ gives

$$\left. \frac{\partial K_d(x, y)}{\partial x} \right|_{x=\Pi^\dagger(x_r)} = d \langle x_r, (T_q^\top K_o T_q)^{-1} \Pi(y) \rangle^{\frac{d-1}{d}} y^\top$$

which immediately gives an expression for $J_\Pi(\Pi^\dagger(x_r))$. Note that this approximation is global in the sense that the $q \times q$ matrix inverse $(T_q^\top K_o T_q)^{-1}$ need only be computed once³; no updating is required during simulation of the closed system.

C. Reduced System Dynamics

Given an estimate $\hat{f}(\Pi(x), u)$ of $f(x, u)$ in the RKHS \mathcal{H}_f and a notion of $J_\Pi(\Pi^\dagger(x_r))$ from above, we can write down a closed dynamical system on the reduced statespace. We have

$$\begin{aligned}
\dot{x}_r &\approx \left(J_\Pi(x) \hat{f}(\Pi(x), u) \right) \Big|_{x=\Pi^\dagger(x_r)} \\
&\approx \left(J_\Pi(x) \right) \Big|_{x=\Pi^\dagger(x_r)} \mathbf{C}^\top \mathbf{k}^f(x_r, u) \\
&\approx T_q^\top J_{\mathbf{k}}(\Pi^\dagger(x_r)) \mathbf{C}^\top \mathbf{k}^f(x_r, u)
\end{aligned} \tag{40}$$

where \mathbf{C} is a matrix with the vectors \mathbf{c}_i as its rows, and $J_{\mathbf{k}}$ is the Jacobian of the empirical feature map defined in Equation (30). Here the expression $J_{\mathbf{k}}(\Pi^\dagger(x_r))$ should be interpreted as notation for either of the Jacobian approximations suggested in Section IV-B.

Equation (40) is seen to give a closed nonlinear control system expressed solely in terms of the reduced variable $x_r \in \mathbb{R}^q$:

$$\begin{cases} \dot{x}_r = T_q^\top J_{\mathbf{k}}(\Pi^\dagger(x_r)) \mathbf{C}^\top \mathbf{k}^f(x_r, u) \\ y = \hat{h}(x_r) \end{cases} \tag{41}$$

where the map $\hat{h} \circ \Pi$ modeling the output function $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ is estimated as described immediately below. Although the “true” reduced system does not actually exist due to non-injectivity of the feature map Φ , in many situations one can expect that the above system will capture the essential input-output behavior of the original system. We leave a precise analysis of the error in the approximations appearing in (40) to future work.

D. Outputs of the Reduced System

Analogous to the case of the dynamics f , we are faced with two possibilities for approximating $y = h(\Pi^\dagger(x_r))$. We can apply a crude Taylor series approximation to estimate Π^\dagger and therefore $h(\Pi^\dagger(x_r))$, or as in Section IV-A, we can estimate a map $(\hat{h} \circ \Pi) : \mathbb{R}^n \rightarrow \mathbb{R}^p$, $x_r \mapsto y$ from the reduced state space to the output space directly, using RKHS methods. Given samples $\{\Pi(x_j), y_j\}_{j=1}^\ell$, each coordinate

³We use the word “inverse” loosely. In practice one would use a numerically stable method, such as an LU-factorization, which can be used to rapidly compute $A^{-1}b$ for fixed A but many different b .

function $(\hat{h}_i)_{i=1}^p$ is given in the familiar form $\hat{h}_i(\Pi(x)) = \sum_{j=1}^{\ell} b_j^i K^h(\Pi(x), \Pi(x_j))$, where K^h is the kernel chosen to define the RKHS, and may be different for each coordinate.

It should be noted that just given the state space reduction map Π , one can immediately compare the output of the system defined by $\hat{h}(x_r)$ to the original system without defining a closed dynamics as above. In fact with Π and \hat{h} one can design a simpler controller which takes as input the reduced state variable x_r , but controls the original system.

E. Structural Properties of the Reduced Order System

In this section, we show that a linearization of the reduced order system preserves the important structural properties of a linearization of the full order system. We leave the study of the structural properties of the reduced nonlinear system for future work.

We first note that the approach introduced in this paper reduces to Moore's approach [19] if the system is linear and one adopts the linear kernel $K(x, y) = \langle x, y \rangle$. For instance, consider the linear control system (1). If we take $\Phi(x) = x$, then the empirical controllability and observability Gramians in feature space are exactly the Gramians (23)-(24), as introduced in Moore's work [19, page 21]. The feature space is the original data space, so balancing in the RKHS as explained above reduces to Moore's notion of balancing. In this case one obtains reduced order system dynamics via a Galerkin projection rather than by attempting to statistically estimate the (linear) dynamics and output functions⁴.

The following brief Proposition shows that the reduced order system obtained using the methods proposed here preserves important properties of the original system.

Proposition 4.1: If the linearization of the full order system (7) is controllable/observable/Hurwitz then the linearization of the reduced order system (41) is controllable/observable/Hurwitz.

Proof: From (37) and (40), the dynamics of the reduced order system is given by

$$\begin{aligned} \dot{x}_r &= (J_{\Pi}(x)f(x, u)) \Big|_{x=\Pi^{\dagger}(x_r)}, \\ &\approx (J_{\Pi}(x)\hat{f}(\Pi(x), u)) \Big|_{x=\Pi^{\dagger}(x_r)} \\ &\approx (J_{\Pi}(x)) \Big|_{x=\Pi^{\dagger}(x_r)} \mathbf{C}^{\top} \mathbf{k}^f(x_r, u) \\ &\approx T_q^{\top} J_{\mathbf{k}}(\Pi^{\dagger}(x_r)) \mathbf{C}^{\top} \mathbf{k}^f(x_r, u) \\ &= A_r x_r + B_r u + O(x_r, u)^2, \end{aligned} \tag{42}$$

and the output is given by

$$y = \hat{h}(x_r) = C_r x_r + O(x_r)^2,$$

where $(A_r, B_r, C_r) \in \mathbb{R}^{q \times q} \times \mathbb{R}^{m \times q} \times \mathbb{R}^{q \times p}$. Because our approach reduces to Moore's approach in the linear case, then using Moore's results in [19] we may conclude that if the linearization of the full order system is asymptotically stable, controllable, and observable (which is the case under assumption H), then A_r is Hurwitz, (A_r, B_r) is controllable and (A_r, C_r) is observable. ■

F. Algorithm Summary

To summarize, the approach we have proposed proceeds as follows

- 1) Given a nonlinear control system (4), let $u^i(t) = \delta(t)e_i$ be the i -th excitation signal for $i = 1, \dots, m$, and let $x^i(t) : t \in [0, \infty) \mapsto x^i(t) \in \mathbb{R}^n$ be the corresponding response of the system. Run the system and sample the trajectories at times $\{t_j\}_{j=1}^N$ to generate a collection of $N \cdot m$ vectors $\{x^i(t_j) \in \mathbb{R}^n\}$.

⁴That is, whenever the system is known. If not, one will still of course need to estimate the system statistically by sampling trajectories and applying the procedure outlined above.

- 2) Fixing $u(t) = 0$ and setting $x_0 = e_i$ for $i = 1, \dots, n$ (separately), measure the corresponding system output responses $y^i(t) : t \in [0, \infty) \mapsto y^i(t) \in \mathbb{R}^p$. As before, sample the responses at times $\{t_j\}_{j=1}^N$ and save the collection of $N \cdot p$ vectors $\{d_k(t_i)\}$ defined as

$$d_k(t_j) = (y_k^1(t_j), \dots, y_k^n(t_j))^\top \in \mathbb{R}^n, \quad k = 1, \dots, p, \quad j = 1, \dots, N \quad (43)$$

- 3) Choose a kernel K defining a RKHS \mathcal{H} , and form the Hankel kernel matrix $K_{o,c} \in \mathbb{R}^{Np \times Nm}$,

$$(K_{o,c})_{\mu\nu} = K(d_\mu, x_\nu) \quad \mu = 1, \dots, Np, \quad \nu = 1, \dots, Nm \quad (44)$$

where we have re-indexed the sets $\{d_k(t_i)\} = \{d_\mu\}$, $\{x^i(t_j)\} = \{x_\nu\}$ to use single indices.

- 4) Compute the eigendecomposition $K_{o,c}K_{o,c}^\top = V\Sigma^2V^\top$ assuming $K_{o,c}$ has been centered according to Equation (29).
 5) The order of the model is reduced by discarding small eigenvalues $\{\Sigma_{ii}\}_{i=q+1}^n$, and projecting onto the subspace associated with the first $q < n$ largest eigenvalues. This leads to the state-space reduction map $\Pi : \mathbb{R}^n \rightarrow \mathbb{R}^q$ given by

$$\Pi(x) = T_q^\top \mathbf{k}_o(x), \quad x \in \mathbb{R}^n \quad (45)$$

where $T_q = V_q \Sigma_q^{-1/2}$ and $\mathbf{k}_o(x)$ is the centered empirical observability feature map given by Equation (31).

- 6) From input/output pairs or simulated/measured trajectories, learn approximations of the dynamics and output function defined on the reduced state space using, for instance, the method described in section §III-A (equations (10)-(11)). The RKHS used to approximate these functions need not be the same as the RKHS in which balanced truncation was carried out.
 7) Approximate the Jacobian contribution as described in section §IV-B.
 8) Combine the approximations to determine an expression for a closed, reduced, nonlinear dynamical system as described in sections §. IV-C and §. IV-D.

V. EXPERIMENTS

We demonstrate an application of our method on two examples appearing in [21] (Examples 3.1. and 3.2, pgs. 52-54).

A. Model Systems

- 1) *Two-Dimensional Exactly Reducible System:* Consider the nonlinear system

$$\begin{aligned} \dot{x}_1 &= -3x_1^3 + x_1^2x_2 + 2x_1x_2^2 - x_2^3 \\ \dot{x}_2 &= 2x_1^3 - 10x_1^2x_2 + 10x_1x_2^2 - 3x_2^3 - u \\ y &= 2x_1 - x_2. \end{aligned}$$

It can be shown that this system has the same input-output relationship as the system $\dot{y} = -y^3 + u$ by rearranging terms so that

$$\begin{aligned} \dot{x}_1 &= -(2x_1 - x_2)^2x_1 + (x_1 - x_2)^3 \\ \dot{x}_2 &= -(2x_1 - x_2)^2x_2 + 2(x_1 - x_2)^3 - u \\ y &= 2x_1 - x_2. \end{aligned}$$

Defining the new variables $z_1 = 2x_1 - x_2$ and $z_2 = x_1 - x_2$, the system can then be re-written

$$\begin{aligned} \dot{z}_1 &= -z_1^3 + u \\ \dot{z}_2 &= -z_1^2z_2 - z_2^3 + u \\ y &= z_1. \end{aligned}$$

It can be seen that the variable z_2 may be truncated because it doesn't appear in the expression of the output and thus doesn't affect z_1 .

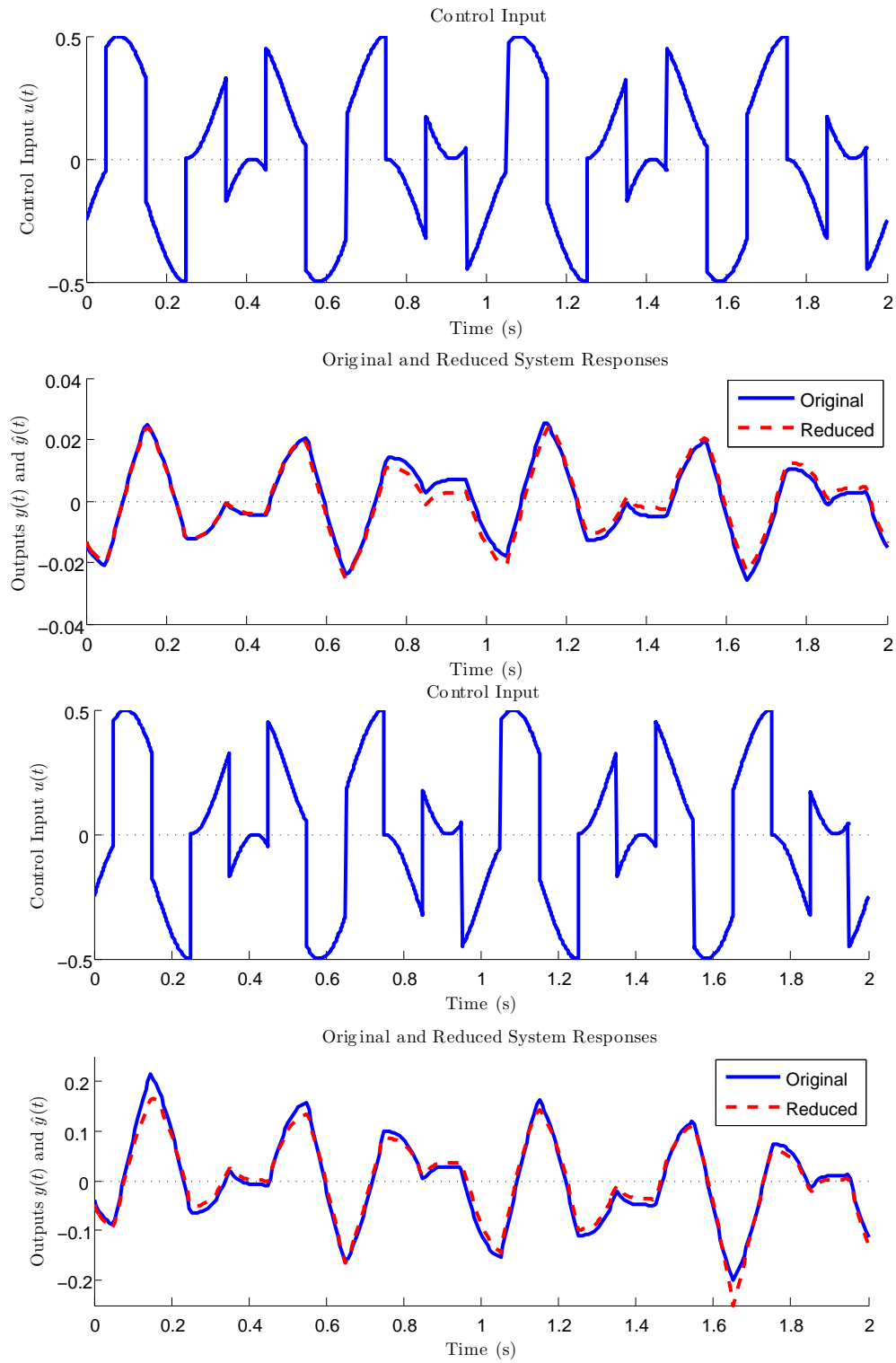


Fig. 1. (Top) Simulated output trajectories for the original and reduced 2-dimensional system. (Bottom) Simulated output trajectories for the original and reduced 7-dimensional system.

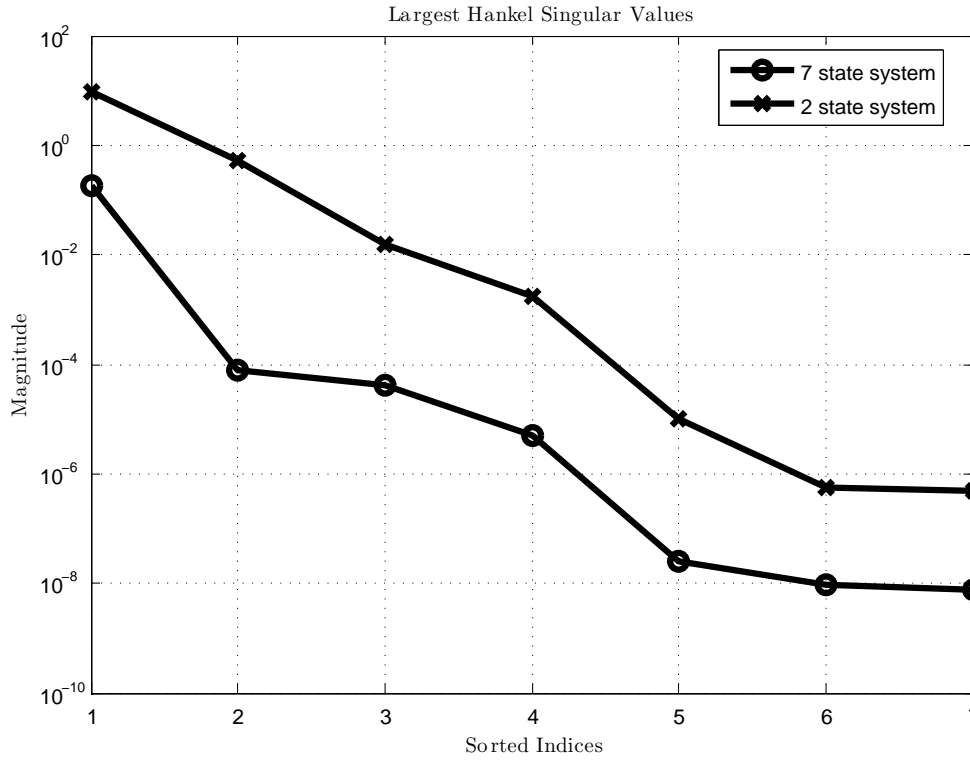


Fig. 2. Largest Hankel singular values of the 2- and 7-dimensional systems in feature space, plotted in descending order. Note that the ordinate axis follows a logarithmic scale.

2) *Seven-Dimensional System:* We will also consider a 7-dimensional nonlinear system with one dimensional input and output:

$$\begin{aligned}
 \dot{x}_1 &= -x_1^3 + u & \dot{x}_2 &= -x_2^3 - x_1^2 x_2 + 3x_1 x_2^2 - u \\
 \dot{x}_3 &= -x_3^3 + x_5 + u & \dot{x}_4 &= -x_4^3 + x_1 - x_2 + x_3 + 2u \\
 \dot{x}_5 &= x_1 x_2 x_3 - x_5^3 + u & \dot{x}_6 &= x_5 - x_6^3 - x_5^3 + 2u \\
 \dot{x}_7 &= -2x_6^3 + 2x_5 - x_7 - x_5^3 + 4u \\
 y &= x_1 - x_2^2 + x_3 + x_4 x_3 + x_5 - 2x_6 + 2x_7
 \end{aligned}$$

B. Experimental Setup

For both systems impulse and initial-condition responses of the system were simulated as described above, and 800 samples equally spaced in the time interval $[0, 5s]$ were sampled to build the Hankel kernel matrix $K_{o,c}$ given by the third degree polynomial kernel $K(x, y) = (1 + \langle x, y \rangle)^3$. For the 2-D system we retained one component, and for the 7-D system we retained two for the sake of variety. Thus the reduction map Π was defined by taking the top one or two eigenvectors (scaled columns of T) corresponding to the largest Hankel singular values, giving a reduced state space of dimension one or two for the 2-D and 7-D systems, respectively.

Next, a map from the reduced variable x_r to \dot{x} was estimated following Section IV-A. The same procedure was followed in both experiments. The control input was chosen to be a 10hz square wave with peaks at ± 1 at 50% duty cycle, and 1000 samples from the simulated system in the interval $[0, 5s]$ were mapped down using Π and then used to solve the RLS regression problems, one for each state variable, again using a third degree polynomial kernel. All initial conditions were set to zero. The desired outputs (dependent variable examples) used to learn \hat{f} were taken to be the true function f evaluated at the samples from the simulated state trajectory. We also added a bias dimension of 1's to the data

to account for an offset, and used a fast leave-one-out cross-validation (LOOCV) computation [23] to select the optimal regularization parameter. Two remarks are in order. The above dynamics can in fact be represented explicitly and exactly in a 3rd degree polynomial RKHS; only monomials up to degree 3 appear in the dynamics. Second, the control input is decoupled from the state. Both of these facts can be used to obtain an improved reduced model, however we did not make use of these special properties and instead applied the simplest version of the techniques described above which assume no special structure.

We followed a similar process to learn the output function $y = \hat{h}(x_r)$ for both systems. Here we used a 10Hz square wave control input (peaks at ± 2 , 50% duty cycle), zero initial conditions and 700 samples in the interval $[0, 5s]$. For this function the Gaussian kernel $K(x, y) = \exp(-\gamma\|x - y\|_2^2)$ was used to demonstrate that our method does not rely on any particular match between the form of the dynamics and the type of kernel. The scale hyperparameter γ was chosen to be the reciprocal of the average squared-distance between the training examples. We again used LOOCV to select the RLS regularization parameter.

Finally, closed systems were simulated as described above using $x_0 = 0$ and a control input different from those used to learn the dynamics and output functions: $u(t) = \frac{1}{4}(\sin(2\pi 3t) + \text{sq}(2\pi 5t - \pi/2))$ where $\text{sq}(\cdot)$ denotes the square wave function. This input is shown at the top of both simulation summaries in Figure 1. The Taylor series approximation for Π was done once, about x_0 , and was not updated further.

C. Results

Figure 2 shows the top seven Hankel singular values in the feature space for the two problems on a log scale. One can see that, for both systems, even a single component ought to capture most of the system's behavior. Further simulations of the 7-D system with a single component showed only a small amount of additional error beyond that the two component system, as one would expect from the decay of that system's Hankel values.

The simulated outputs $\hat{y}(t)$ of the closed reduced systems as well as the output $y(t)$ of the original system are plotted in Figure 1 (left, 2-D system ; right, 7-D system). One can see that, even for a significantly different input, the reduced systems closely capture the original systems. The main source of error is seen to be over- and under-shoot near the square wave transients. This error can be further reduced by simulating the system for different sorts of inputs (and/or frequencies) and including the collected samples in the training sets used to learn Π , \hat{f} and \hat{h} . Indeed, we have had some success driving example systems with random uniform input in some cases.

Finally, for illustrative purposes we show examples of the controllability and observability kernel matrices K_c and K_o for the 7-D system in Figure 3.

VI. CONCLUSION

We have introduced a new, empirical model reduction method for nonlinear control systems. The method assumes that the nonlinear system is approximately linear in a high dimensional feature space, and carries out linear balanced truncation in that space. This leads to a nonlinear reduction map, which we suggest can be combined with representations of the dynamics and output functions by elements of an RKHS to give a closed reduced order dynamical system which captures the input-output characteristics of the original system. We then demonstrated an application of our technique to a pair of nonlinear systems and simulated the original and reduced models for comparison, showing that the approach proposed here can yield good low-order nonlinear reductions of strongly nonlinear control systems. We believe that techniques well known to the machine learning and statistics communities can offer much to control and dynamical systems research, and many further directions remain, including computing error estimates, reduction of unstable systems, structure preserving systems, stochastically perturbed systems, and finding easily verifiable conditions of model reducibility of nonlinear systems.

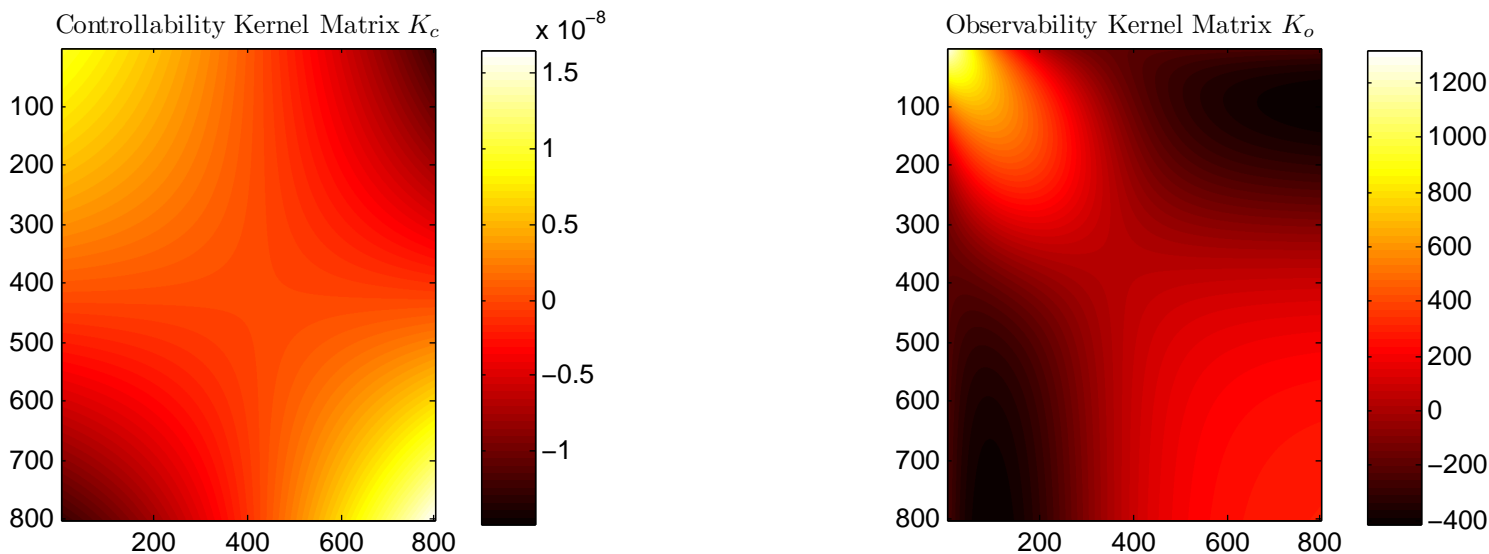


Fig. 3. Plots of the kernel matrices encoding controllability properties (left) and observability properties (right) of the 7-dimensional system.

REFERENCES

- [1] Antoulas, A. C. (2005). *Approximation of Large-Scale Dynamical Systems*, SIAM Publications.
- [2] Aronszajn, N. (1950). Theory of Reproducing Kernels. *Trans. Amer. Math. Soc.*, 68:337-404.
- [3] Bouvrie, J. and B. Hamzi (2010). Balanced Reduction of Nonlinear Control Systems in Reproducing Kernel Hilbert Space, Proc. 48th Annual Allerton Conference on Communication, Control, and Computing, 2010, pp. 294 – 301.
- [4] Coifman, R. R., I. G. Kevrekidis, S. Lafon, M. Maggioni and B. Nadler (2008). Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Systems, *Multiscale Model. Simul.*, 7(2):842-864.
- [5] Cucker, F. and S. Smale (2001). On the mathematical foundations of learning. *Bulletin of AMS*, 39:1-49.
- [6] Dullerud, G. E., and F. Paganini (2000). *A Course in Robust Control Theory: a Convex Approach*, Springer.
- [7] Fujimoto, K. and D. Tsubakino (2008). Computation of nonlinear balanced realization and model reduction based on Taylor series expansion, *Systems and Control Letters*, 57, 4, pp. 283-289.
- [8] Gray, W. S. and E. I. Verriest (2006). Algebraically Defined Gramians for Nonlinear Systems, *Proc. of the 45th IEEE CDC*.
- [9] Jonckheere, E.A., and L. M. Silverman (1983). A New Set of Invariants for Linear Systems - Application to Reduced Order Compensator Design. *IEEE Transactions on Automatic Control*, AC-28, 10, 953-964.
- [10] Kenney, C. and G. Hewer (1987). Necessary and Sufficient Conditions for Balancing Unstable Systems, *IEEE Transactions on Automatic Control*, 32, 2, pp. 157-160.
- [11] Krener, A. (2006). Model reduction for linear and nonlinear control systems. Bode Lecture, 45th IEEE Conference on Decision and Control.
- [12] Krener, A. J. (2007). The Important State Coordinates of a Nonlinear System. In “*Advances in control theory and applications*”, C. Bonivento, A. Isidori, L. Marconi, C. Rossi, editors, pp. 161-170. Springer.
- [13] Krener, A. J. (2008). Reduced order modeling of nonlinear control systems. In “*Analysis and Design of Nonlinear Control Systems*”, A. Astolfi and L. Marconi, editors, pp. 41-62. Springer.
- [14] Kwok, J. T. and I.W. Tsang (2003). “The Pre-Image Problem in Kernel Methods”. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*.
- [15] Lall, S., J. Marsden, and S. Glavaski (2002). A subspace approach to balanced truncation for model reduction of nonlinear control systems, *International Journal on Robust and Nonlinear Control*, 12, 5, pp. 519-535.
- [16] Laub, A.J. (1980). On Computing “balancing” transformations, *Proc. of the 1980 Joint Automatic Control Conference (ACC)*.
- [17] Li, J.-R. (2000). *Model Reduction of Large Linear Systems via Low Rank System Grammians*. Ph.D. thesis, Massachusetts Institute of Technology.
- [18] Mika, S., B. Schölkopf, A. Smola, K. R. Müller, M. Scholz, and G. Rätsch (1998). Kernel PCA and de-noising in feature spaces, In *Proc. Advances in Neural Information Processing Systems (NIPS) 11*, pp. 536–542, MIT Press.
- [19] Moore, B. (1981). Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction, *IEEE Tran. Automat. Control*, 26, 1, pp. 17-32.
- [20] Newman, A.J., and P. S. Krishnaprasad (2000). Computing balanced realizations for nonlinear systems, *Proc. of the Math. Theory of Networks and Systems (MTNS)*.
- [21] Nilsson, O. (2009). *On Modeling and Nonlinear Model Reduction in Automotive Systems*, Ph.D. thesis, Lund University.
- [22] Phillips, J., J. Afonso, A. Oliveira and L. M. Silveira (2003). Analog Macromodeling using Kernel Methods. In *Proceedings of the IEEE/ACM International Conference on Computer-aided Design*.
- [23] Rifkin, R., and R.A. Lippert. *Notes on Regularized Least-Squares*, CBCL Paper 268/AI Technical Report 2007-019, Massachusetts Institute of Technology, Cambridge, MA, May, 2007.
- [24] Scherpen, J. (1994). *Balancing for Nonlinear Systems*, Ph.D. thesis, University of Twente.

- [25] Schölkopf, B., Smola, A., and Müller, K (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.
- [26] Therapos, C. P. (1989). *Balancing Transformations for Unstable Nonminimal Linear Systems*, IEEE Transactions on Automatic Control, **34**, 4, pp. 455-457.
- [27] Wahba, G. (1990). Spline Models for Observational Data, *SIAM CBMS-NSF Regional Conference Series in Applied Mathematics* 59.
- [28] Weiland, S. (1991). Theory of approximation and disturbance attenuation of linear systems, Doctoral dissertation, University of Groningen.